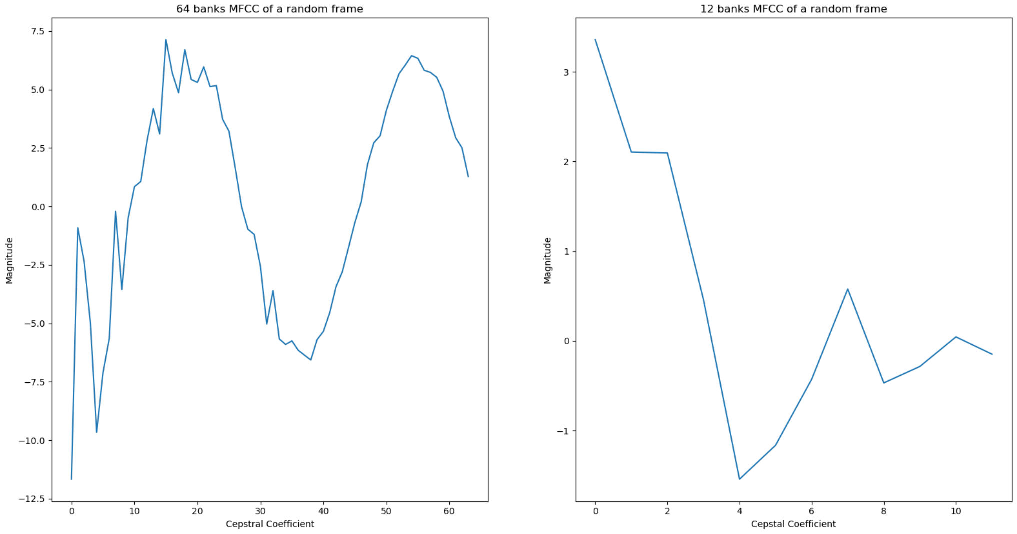
DSP Lab02

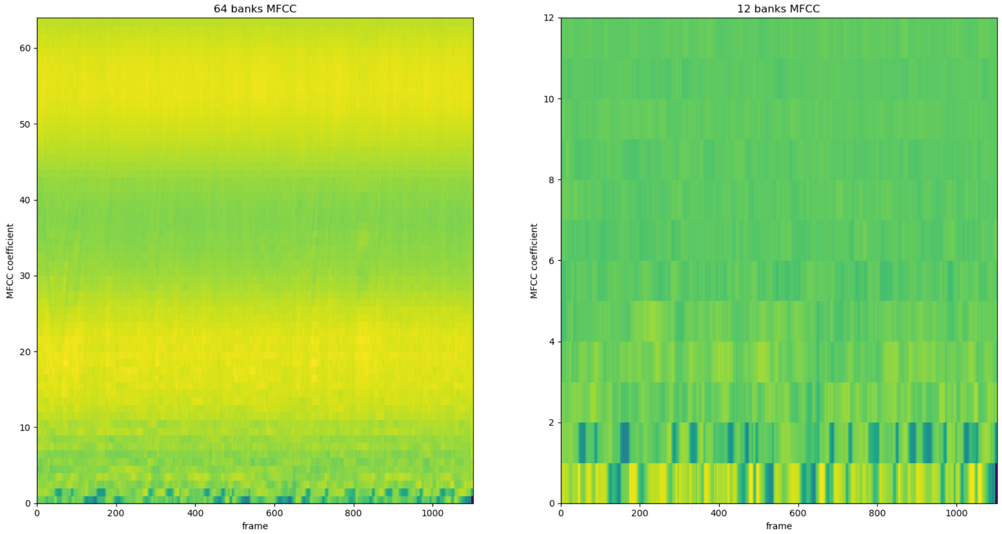
Audio & Speech: Audio Reconstruction from MFCC

108061129 陳楷芮

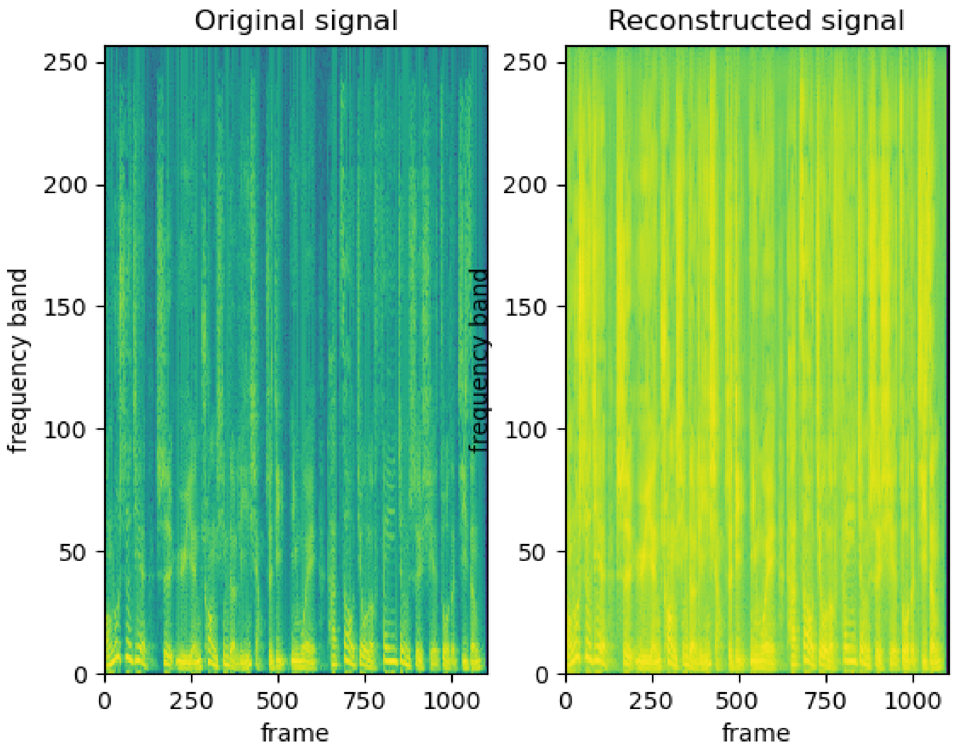
Demo(1): **Effect of num Mel-Filter banks**



Demo(2): **Effect of num Mel-filter banks**



Demo(3): **Ori vs Reconstructed**



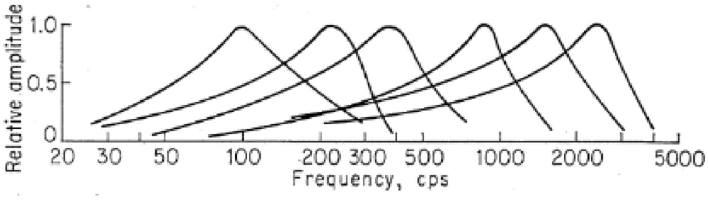
Demo(4): **64-banks pre-emphasized reconstructed audio**

uploaded to eeclass

Handout Questions

1. Why not using rectangular filters for “energy” calculation?

The reason of using triangular filters is because of the structure of human ear. Cochlea has different frequency responses to pure tones in different parts of itself. The frequency-amplitude distribution tends to behave as below.



von Békésy (1943)

It’s quite obvious that the behavior looks like several triangular filters with its corresponding resonant frequencies. Therefore, to imitate the function of cochlea, we use triangular filters in energy calculation.
Rectangular filters can work as well though, but triangular is much better, since it’s closer to what human ear hears.

2. Why should mel-filters be overlapping?

It’s because we don’t want to lose energy. If mel-filters don’t overlap, when performing convolution, some frequencies would be reduced and lead to total energy loss.

3. Why are high-quefrency MFCCs usually abandoned when doing speech recognition?

Since human ear is more sensitive to low-frequencies(as we can see from the figure above that shows that the bandwidth of high frequency is larger), therefore high-frequency part is usually discarded.

4. Do you think MFCC is good for speaker identification purposes

Yes. Since different speakers have different formant information which leads to recognizable MFCC features, therefore we can use these features for speaker recognition.

4.1 How about baby-cry detection?

For MFCC, since it has lower resolution in high-frequencies, it may not be suitable for baby-cry detection, which mostly consists of high-frequency sounds.

5. If two sounds have similar MFCCs, does that imply they sound similar to our ears?

Yes, since mel-scale imitates what our cochlea functions, which has higher resolution in low frequencies and lower resolution in high frequencies, making MFCC more similar to what our human ear perceives.

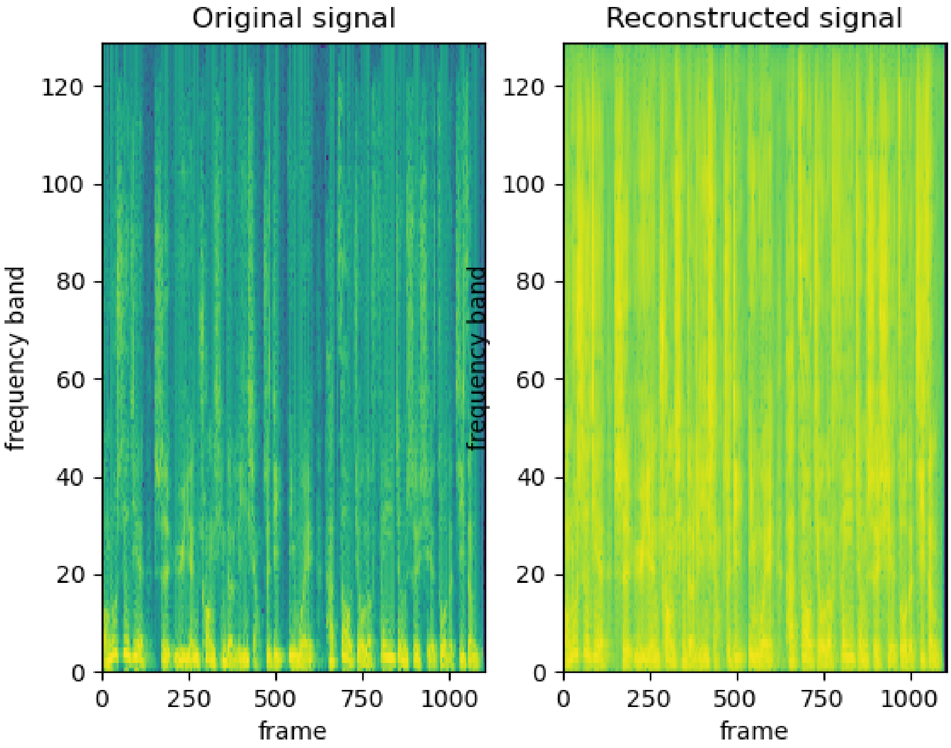
Report Questions

1. What are the artifacts and distortions in the reconstructed audio? Suggest what the causes of these degradations are.

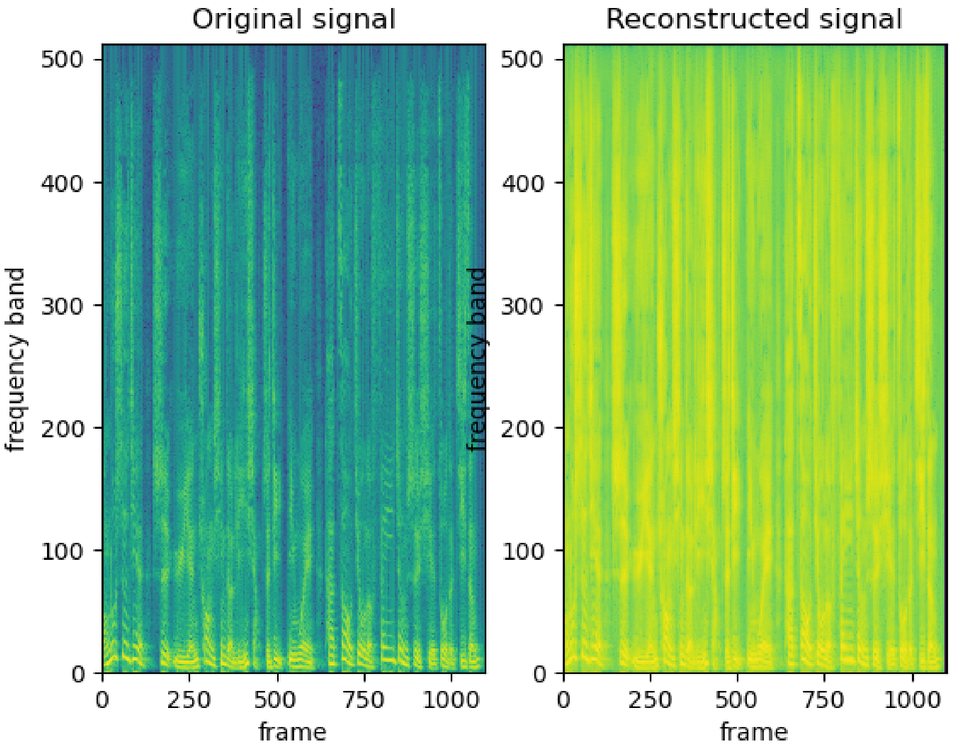
Since in the MFCC process, phase is discarded, therefore causing the distortion in the rconstructed audio. Even when we apply the griffin-lim algorithm to reconstruct the phase, it still has some bias between the real one and the estimated one.

2. Experiment with different frame length, step length, and the number of fbanks; discuss what effects each of them has in the reconstruction process.

1. Frame length discussion
    a. Frame length = 128

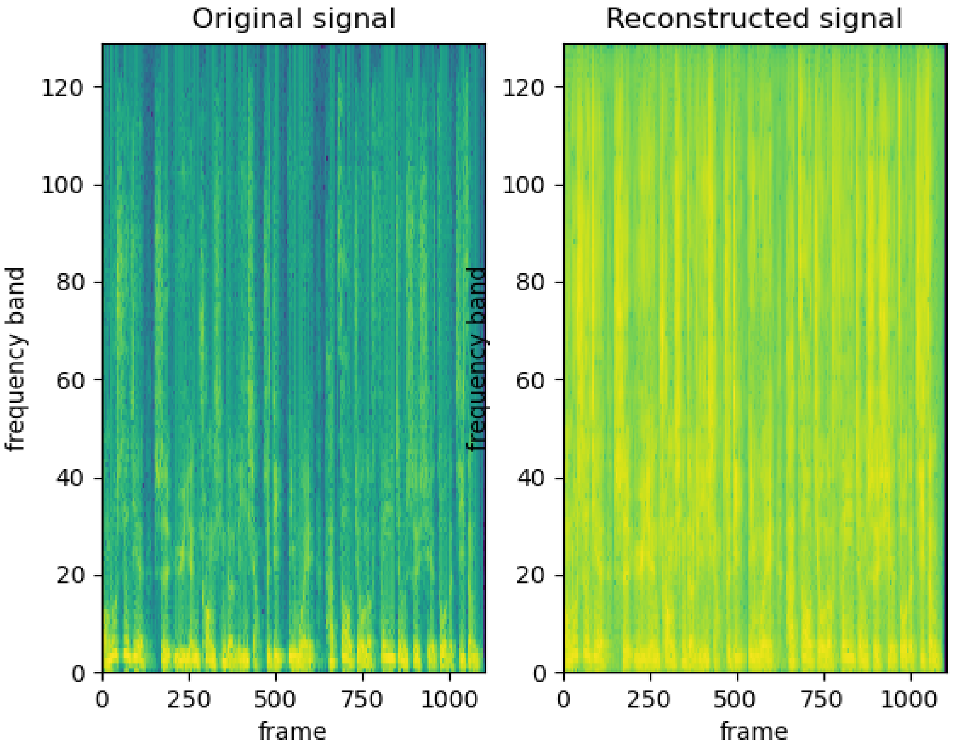


b. Frame length = 1024

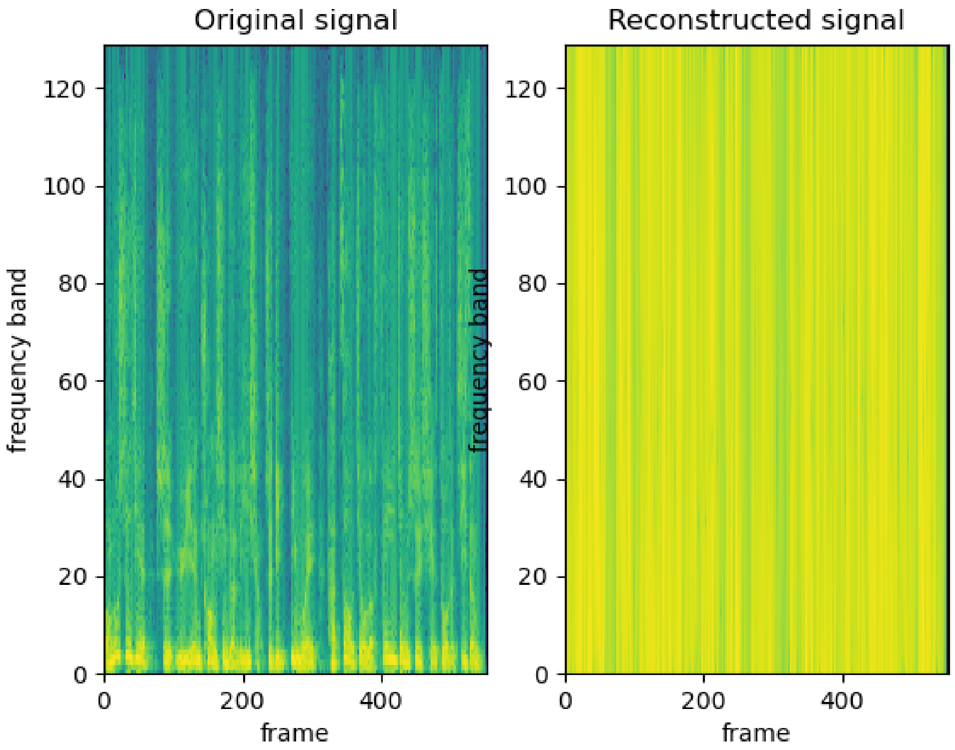


As we can see from the above plots, although the difference is hardly discoverable, we can still see that frame length 1024 has higher resolution than frame length 128.

2. step length discussion
   a. Step length = 128



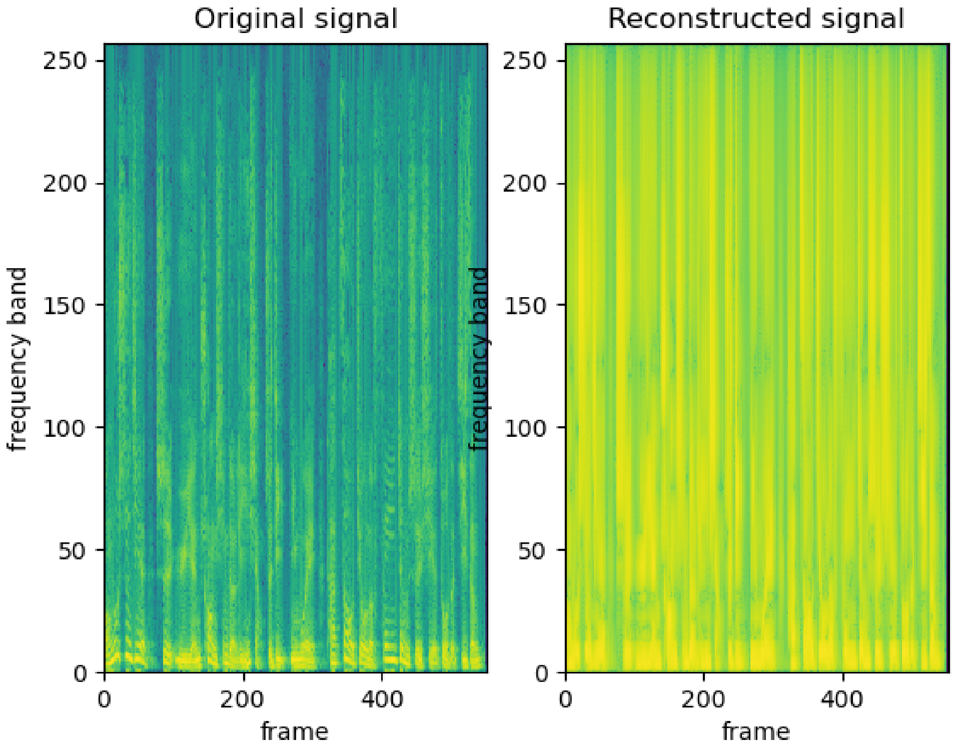
   b. Step length = 256



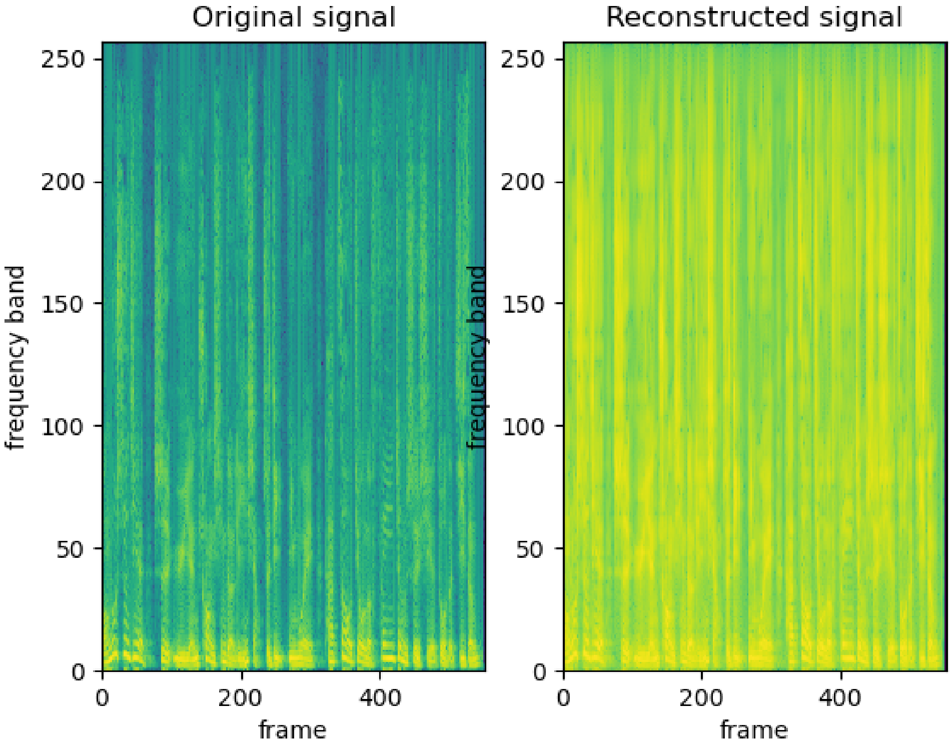
As we can see from the above plots, step length larger meaning that more signals are loss, causing the number of frame smaller.

3. Number of fbanks

   a. number of fbanks: 12



   b. number of fbanks: 64



As we can see from the plots, when the number of fbanks is larger, meaning that the MFCC contains more information of the original signal.